Применение BERT к задаче автоматической категоризации постов

Выпускная квалификационная работа

Кизеев Данил Владимирович
Научный руководитель: Липкович Михаил Маркович
kizeevdanil@yandex.ru

СПбГУ

29 июня 2022 г.

Постановка задачи

Даны:

- Категории $C = [$мода, фитнесс, гармония, видеоигры, красота, животные, юмор, искусство, путешествия, музыка, танцы, спорт, окружающая среда, техника, еда$]$.
- Множество публикаций (изображений либо видео) в социальной сети, называемыми «постами». Обозначим данное множество как L.
- Распределения постов по меткам.

Требуется: каждой категории из множества C поставить в соответствие наиболее подходящие объекты из множества L.

Пример представления данных

Рис. 1: Пример объекта: изображение мужчины.

```
('images/c_f_51bcd1a20cf23bf9736ecd58db607b105982e2.jpeg',
 [{'description': 'Pool player', 'score': 0.987},
  {'description': 'Indoor games and sports', 'score': 0.9811},
  {'description': 'Cue stick', 'score': 0.9795},
  {'description': 'Pool', 'score': 0.9768},
  {'description': 'Recreation room', 'score': 0.9736},
  {'description': 'Recreation', 'score': 0.972},
  {'description': 'Billiard room', 'score': 0.9657},
  {'description': 'Shoulder', 'score': 0.9656},
  {'description': 'Textile', 'score': 0.9583},
  {'description': 'Joint', 'score': 0.954}])
```

Рис. 2: Распределение по меткам.

Три модели решения данной задачи

Для решения данной задачи предлагаются три метода:

- Используя векторные представления слов BERT[1].
- Используя модель BERTopic[2] для нахождения тематик в документах.
- Используя векторные представления слов word2vec[3] и статистическую меру важности слов в документах TF-IDF[4].

[1] **Devlin J. et al.** *Bert: Pre-training of deep bidirectional transformers for language understanding.* //arXiv preprint arXiv:1810.04805. – 2018.

[2] **Maarten Grootendorst** *BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics.* //Zenodo (2020) https://doi.org/10.5281/zenodo.4381785

[3] **Mikolov T. et al.** *Efficient estimation of word representations in vector space.* //arXiv preprint arXiv:1301.3781. – 2013.

[4] **Juan Ramos J. et al.** *Using tf-idf to determine word relevance in document queries* //Proceedings of the first instructional conference on machine learning. – 2003. – Т. 242. – №. 1. – С. 29-48.

Модель BERT

Принцип работы предложенного алгоритма

1. Составление предложений из слов. Производится простой склейкой.
2. Представление предложений в векторном виде, используя предобученную нейросеть **BERT**.
3. Понижение размерности полученных векторов с помощью **UMAP**[a].
4. Кластеризация векторов объектов, используя метод **средних соседей**. Оптимальное количество находим по метрике **среднего силуэта**.
5. Возвращаемся в исходную размерность и составляем кластеры в соответствии с кластерами для пониженной размерности.
6. Сопоставление каждой категории постов, используя **косинусную близость**.
7. Для оценки качества используются **матрица спутанностей**[b] и метрики F-меры, полноты и точности.

[a] **Leland McInnes, John Healy, James Melville** *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* //arXiv, 2018. — 64 стр.

[b] **Stehman S. V.** *Selecting and interpreting measures of thematic classification accuracy.* //Remote sensing of Environment. – 1997. – Т. 62. – №. 1. – С. 77-89.

Проблема с расстояниями между векторами

Чтобы добиться сравнения векторов категорий с векторами меток, нужно, чтобы векторы имели одинаковую размерность. При снижении числа компонент для малого количества векторов (в нашем случае категорий 15 штук) пропадают исходные зависимости, а именно: косинусы углов изменяются и хорошего качества добиться трудно.

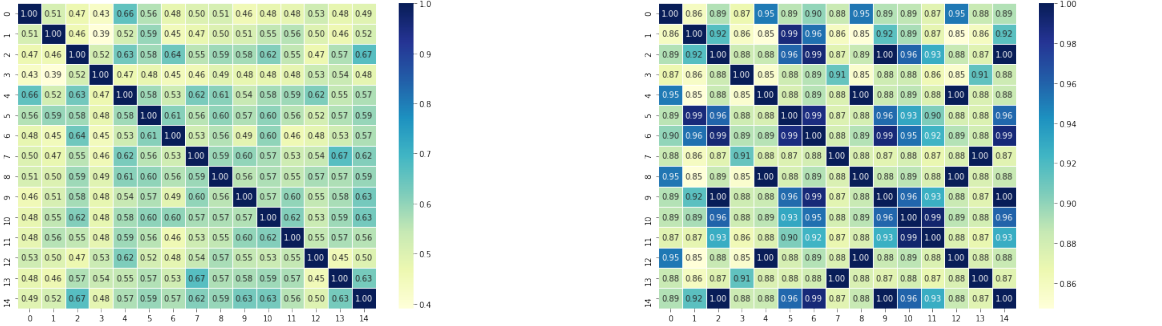

Рис. 3: Видно, что зависимости между расстояниями не сохраняются.

Оценка качества первой модели

Для оценки качества будем использовать три метрики — F-меру, точность и полноту.

Также будем смотреть на матрицу ошибок — это матрица различий между эталонными (размеченными) и полученными моделью значениями.

Значение F-меры равно 0.08 на наших данных. Из-за малого количества рекомендаций (всего 3000 из 36000) признать эту модель годной нельзя.

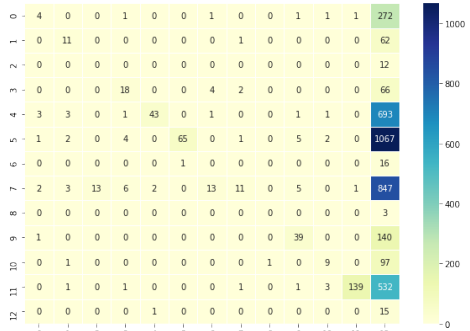

Рис. 4: Матрица спутанностей для первой модели.

BERTopic

Принцип работы второй модели

1. Составление предложений из слов.
2. Представление полученных предложений в векторном виде (BERT).
3. Обучение модели **c-TF-IDF**[a]. Отличие от стандартной модели в объединении документов из одного кластера в один.
4. Понижение размерности для векторов (UMAP).
5. Кластеризация векторов объектов в векторном пространстве методом **HDBSCAN**[b]. Все выбросы удаляются — в нашем случае это 1/3 всего датасета.
6. Сопоставление каждой категории постов, используя косинусную близость. Найдём оптимальный порог «доверия» нашей модели.

---

[a] **Özgür A., Özgür L., Güngör T.** *Text categorization with class-based and corpus-based keyword selection* //International Symposium on Computer and Information Sciences. – Springer, Berlin, Heidelberg, 2005. – С. 606-615.

[b] **Campello R.J.G.B., Moulavi D., Sander J.** *Density-Based Clustering Based on Hierarchical Density Estimates*. //Lecture Notes in Computer Science, vol 7819. Springer, Berlin, Heidelberg, 2013.

Оценка качества

Используем матрицу спутанностей для визуальной оценки качества. Также найдём оптимальное значение порога для лучшей обобщающей способности.

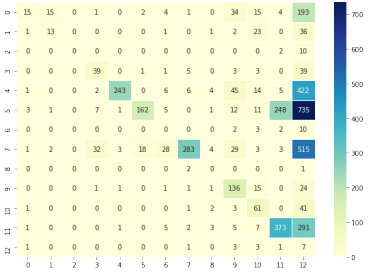

Рис. 5: Матрица спутанностей для вектора предсказаний. Уже визуально можно сказать, что модель точна, но скудна на количество предсказаний.

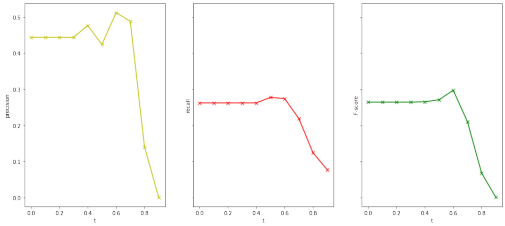

Рис. 6: Значения метрик точности, полноты и F-меры для разных значений порога $t = 0, 0.1, ..., 0.9$.

Оптимальное значение порога равно 0.6 со значением F-меры, равным 0.3.

Третья модель: word2vec

Для лучших предсказаний векторы категорий были получены аггрегацией размеченных объектов вручную.

Принцип работы третьей модели

1. Представление слов в векторном виде (**word2vec**).
2. Обучение модели TF-IDF и нахождение весов слов. Поскольку некоторые метки состоят из нескольких слов, то обучения производится по n-граммам от 1 до 3.
3. Сопоставление каждой категории постов по косинусу угла между двумя векторами. Нахождение порога для оптимальной обобщающей способности.

Качество модели word2vec

Используем матрицу спутанностей для визуальной оценки качества. Также найдём оптимальное значение порога на значения косинусов углов между векторами для нахождения оптимальной обобщающей способности.

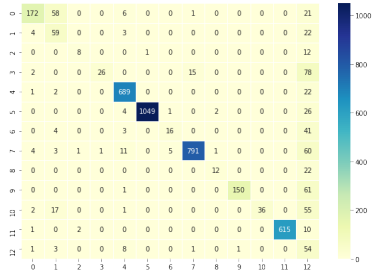

Рис. 7: Матрица спутанностей для вектора предсказаний третьей модели.

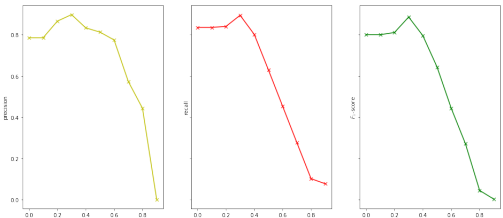

Рис. 8: Значения метрик точности, полноты и F-меры для разных значений порога $t = 0, 0.1, ..., 0.9$.

Оптимальное значение порога равно 0.3 со значением F-меры, равным 0.89.

Итоги

- Для решения задачи категоризации при построении моделей были успешно использованы методы машинного обучения и нейронных сетей.
- Были предложены три модели, в порядке увеличения качества предсказаний.
- Первая модель показала малое количество предсказаний и плохое качество. С помощью второй модели удалось улучшить как точность, так и количество предсказаний. Наконец, в третьей модели был показан значительный прирост качества предсказаний.

Спасибо за внимание